Synthetic lethality-based prediction of cancer treatment response from histopathology images

Danh-Tai Hoang[1], Gal Dinstag[2], Leandro C. Hermida[3], Doreen S. Ben-Zvi[2], Efrat Elis[2], Katherine Caley[1], Sanju Sinha[3], Neelam Sinha[3], Christopher H. Dampier[4], Tuvik Beker[2], Kenneth Aldape[4], Ranit Aharonov[2], Eric A. Stone[1,*], Eytan Ruppin[3,*].

[1] Biological Data Science Institute, College of Science, Australian National University, Canberra, ACT, Australia
[2] Pangea Biomed Ltd., Tel Aviv 6971003, Israel
[3] Cancer Data Science Laboratory, Center for Cancer Research, National Cancer Institute, Bethesda, MD, USA
[4] Laboratory of Pathology, Center for Cancer Research, National Cancer Institute, Bethesda, MD, USA.
* Corresponding authors: eric.stone@anu.edu.au (E.A.S.) and eytan.ruppin@nih.gov (E.R.)

ABSTRACT

Advances in artificial intelligence have paved the way for predicting cancer patients' survival and response to treatment from hematoxylin and eosin (H&E)-stained tumor slides. Extant approaches do so either directly from the H&E images or via prediction of actionable mutations and gene fusions. Here we present the first genetic interactions (GI)-based approach for predicting patient response to treatment, founded on two conceptual steps: (1) First, we build DeepPT, a deep-learning framework that predicts tumor gene expression from H&E slides, and subsequently, (2) we apply ENLIGHT - a previously published GI based approach - to predict patient treatment response from the inferred tumor expression. DeepPT was trained on images and corresponding transcriptomics data of TCGA breast, kidney, lung, and brain tumor samples. Testing DeepPT transcriptomics prediction ability, we find that it generalizes well to predicting the expression of two breast and brain cancer unseen independent datasets. Studying samples from a recently published large multi-omics breast cancer clinical trial, we applied ENLIGHT to the expression predicted by DeepPT from the tumor slides. We find that it successfully predicts true responders with a clinically meaningful hazard ratio of about six. These results put forward a general framework for predicting patient response to a broad array of targeted and checkpoint therapies from the histological images. If corroborated further, the new approach could augment the feasibility of precision oncology in developing countries and in other situations where comprehensive molecular profiling is not available.

INTRODUCTION

Histopathology has long been considered the gold standard of clinical diagnosis and prognosis in cancer. In recent years, molecular markers including tumor gene expression have proven increasingly valuable for enhancing diagnosis and precision oncology. Digital histopathology promises to combine these complementary sources of information using machine learning, artificial intelligence and big data. Key advances are already underway, as whole slide images (WSI) of tissue stained with hematoxylin and eosin (H&E) have been used to computationally diagnose tumors [1–3], classify cancer types [3–8], distinguish tumors with low or high mutation burden [9], identify genetic mutations [2,10–17], predict patient survival [18–22], detect DNA methylation patterns [23] and mitoses [24], and quantify tumor immune infiltration [25]. Moreover, the ability to infer gene expression from WSI has also been explored [26–30]. These impressive advances are unravelling the potential of harnessing next-generation digital pathology to predict a patient's response to therapies directly from images [31,32]. The current study aims to take these efforts one step further, demonstrating, for the first time, the feasibility of tumor profiling based on transcriptomic imputation from H&E slides.

To realize this goal, we have taken a two-step approach. First, we developed DeepPT (Deep Pathology for Treatment), a novel deep-learning framework for predicting gene expression from H&E slides. Second, utilizing the predicted gene expression, we apply our previously published approach, ENLIGHT [33], which has originally been developed to predict patients response from the measured tumor gene expression, to predict patients' response from the DeepPT *predicted* transcriptomics.

We proceed to provide an overview of DeepPT architecture and a brief recap of ENLIGHT workings, the study design and the cohorts analysed. We then describe the results obtained, showing the success of the trained DeepPT models in predicting the expression in four TCGA cohorts and two independent cohorts. The crux of our results is focused on applying DeepPT together with ENLIGHT to recently published breast cancer clinical trial data, where it successfully predicts the true responders among breast cancer patients directly from the H&E images, obtaining a clinically relevant hazard ratio of about six. Overall, our results show for the first time that combining digital pathology with expression-based approaches offers an exciting new way to extend the feasibility of precision oncology to the realm of developing countries and to other situations where tumor sequencing is not feasible.

RESULTS

The computational pipeline

DeepPT is trained on formalin-fixed, paraffin-embedded (FFPE) whole slide images and their corresponding gene expression profiles from TCGA patient samples. The model obtained is then used to predict gene expression from both internal held-out and external datasets. In contrast to previous studies aimed at predicting gene expression from WSI, which have focused on fine tuning the last layer of a pre-trained convolutional neural networks (CNN), DeepPT is composed of three main components (**Methods, Figure 1a, and Extended Figure 1**): a CNN model for feature extraction, an auto-encoder for feature compression, and a multiple-layer perceptron (MLP) for the final regression. Intuitively, the pre-trained CNN layers (trained with natural images from ImageNet database [34,35]) play the role of a layperson's eyes that capture the shape and color of images, while the auto-encoder component is reminiscent of the pathologists' expertise in concentrating on the most important histological features or their combinations. The MLP regression module consists of three fully connected layers in which the weights from input layer to hidden layer are shared among genes, enabling the model to capture shared signal among similar gene expression profiles to benefit from the advantage of multi-task learning. Rather than training a model for each gene separately or for all genes together as was done in previous studies [27,29], we trained simultaneously on tranches of genes with similar median gene expression values, allowing shared signal to be leveraged while preventing the model from focusing on only the most highly expressed genes. Overall, DeepPT achieves a marked increase in the accuracy and efficiency compared to current published methods for prediction of gene expression.

The predicted expression then serves as input to ENLIGHT [33], which is a transcriptomics-based approach that predicts individual responses to a wide range of targeted and immunotherapies based on gene expression data measured from the tumor biopsy (**Figure 1b**). ENLIGHT aims to advance and extend the scope of SELECT [36], two recent approaches that rely on analysis of functional genetic interactions (GI) around the target genes of the chosen therapy. Specifically, two broad types of interactions are considered: Synthetic Lethality (SL), whereby the simultaneous loss of two non-essential genes is lethal to the cell, and Synthetic Rescue (SR), whereby the loss of an essential gene can be compensated for through the over- or under-expression of a second gene (its "rescuer" gene). ENLIGHT's drug response

prediction pipeline comprises two steps (**Figure 1b**): (i) Given a drug, the *inference engine* identifies the clinically relevant genetic interaction partners of the drug's target gene(s). The inference engine first identifies a list of initial candidate SL/SR by analysing cancer cell line dependencies based on the principle that SL/SR interactions should decrease/increase tumor cell viability, respectively, when 'activated' (e.g., in the SL case, viability is decreased when both genes are under-expressed). It then selects those pairs that are more likely to be clinically relevant by analysing a database of tumor samples with associated transcriptomics and survival data, requiring a significant association between the joint inactivation of target and partner genes and better patient survival for SL interactions, and analogously for SR interactions. (ii) The drug-specific GI partners are then used to *predict* a given patient's response to the drug based on the gene expression profile of the patient's tumor. The ENLIGHT Matching Score (EMS), which evaluates the match between patient and treatment, is based on the overall activation state of the set of GI partner genes of the drug targets, reflecting the notion that a tumor would be more susceptible to a drug that induces more active SL interactions and fewer active SR interactions.

Here, we predict patient treatment response by applying ENLIGHT to the expression values *predicted* by DeepPT instead of those measured using RNA sequencing. We show that combining ENLIGHT with DeepPT enables robust prediction of response to treatment, in an independent test set of breast cancer patients, for which H&E slides of fresh frozen (FF) samples are available. Remarkably, this is done without adapting either DeepPT, which was trained on FFPE samples from TCGA, or ENLIGHT, which is an unsupervised algorithm, never trained on any response data.

Study design and patient cohorts

The workflow describing the computational analysis is depicted in **Figure 1c.** We collected FFPE WSI together with matched RNAseq gene expression profiles for four major cancer types, including breast, kidney, lung, and brain from TCGA database. Low quality slides were excluded, resulting in 4,368 slides from 3,750 patients. Each cancer type cohort was processed, trained and evaluated separately. We performed a 5×5 nested cross-validation to assess the model performance: In the outer loop, the samples were randomly split into five disjoint sets. Each sample set was selected in turn as the *held-out test set* (20%), while the rest were used for training (80%). Given an outer split, in the inner loop each model was trained five times by

further splitting the training set into internal *training* and *validation* sets, performing a five-fold cross validation. Applying a bagging technique, we averaged the predictions from the five different models, presenting our final prediction for each gene on the held-out test set. The outer loop was repeated five times across the five held-out test sets, hence resulting overall in 25 trained models (**Extended Figure 2a-b**). For further validation, we applied the trained models (with the TCGA cohort) to predict gene expression on two independent external datasets: the TransNEO breast cancer cohort (TransNEO-Breast) consisting of 160 FF slides [37] and a new unpublished brain cancer cohort (NCI-Brain) consisting of 210 FFPE slides, both also containing matched expression data (see **Methods**). Our final goal, however, is to use the predicted gene expression to predict treatment response. To this end, the predicted gene expression served as input to ENLIGHT to predict treatment outcome for patients from TransNEO cohort, as described in the next section.

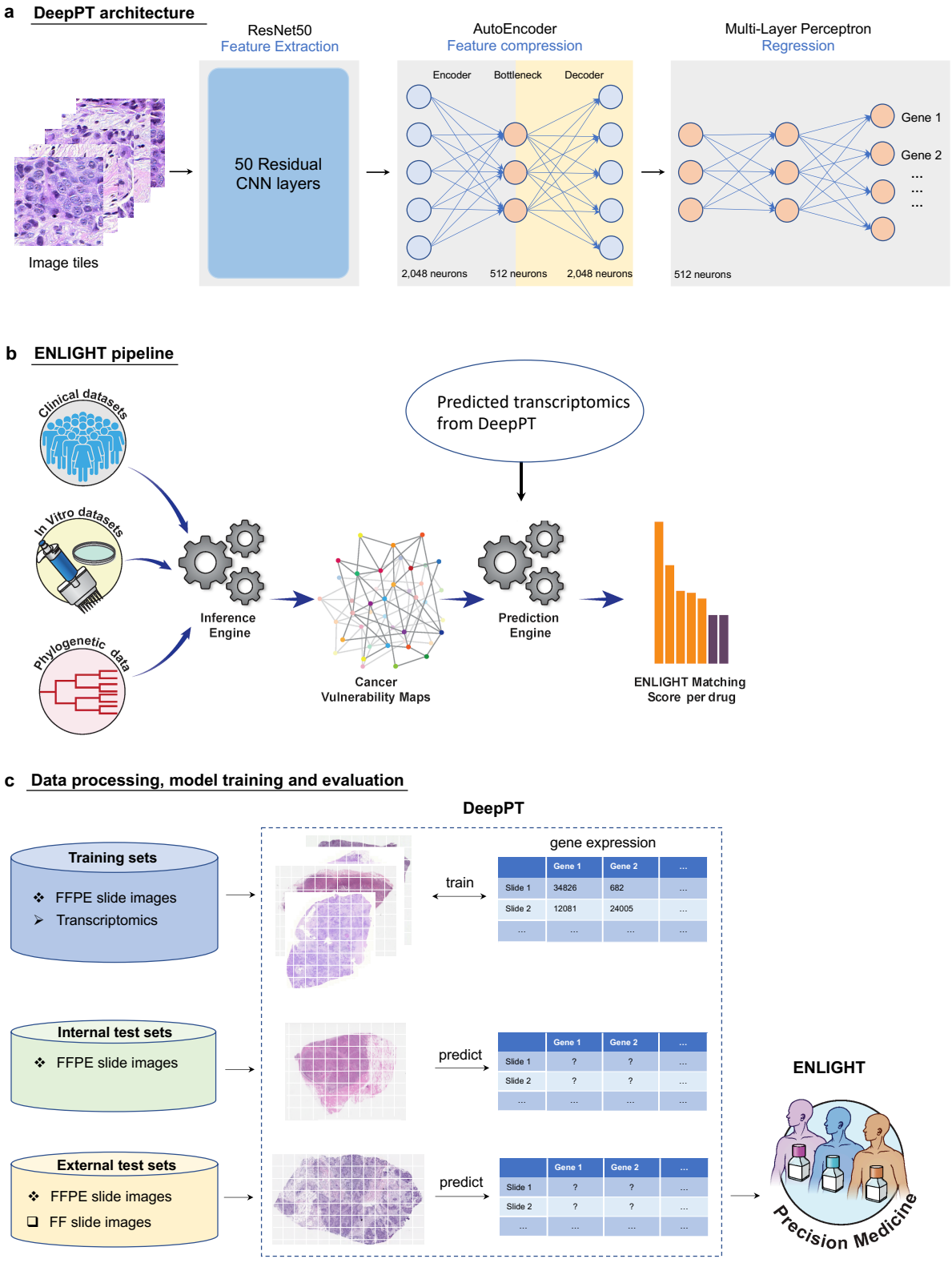


Figure 1. Study overview. (a) The three main components of DeepPT architecture, from left to right. The pre-trained ResNet50 CNN unit extracts histopathology features from tile images. The autoencoder compresses the 2,048 features to a lower dimension of 512 features. The multi-layer perceptron integrates these histopathology features to predict the sample's gene expression. **(b)** An overview of the ENLIGHT pipeline (illustration adapted from [33]:

ENLIGHT starts by inferring the genetic interaction partners of a given drug from various cancer in-vitro and clinical data sources. Given the SL and SR partners and the transcriptomics for a given patient sample, ENLIGHT computes a drug matching score that is used to predict the patient response. Here, ENLIGHT uses DeepPT predicted expression to produce drug matching scores for each patient studied. **(c)** DeepPT was trained with formalin-fixed paraffin-embedded (FFPE) slide images and their matched transcriptomics of TCGA patients from four cancer types, including breast, kidney, lung, and brain. After the training phase, the models were then applied to predict gene expression on the four internal (held-out) TCGA datasets and on two independent datasets on which it was never trained. The predicted tumour transcriptomics served as input to ENLIGHT for predicting patient's response to treatment.

Prediction of gene expression from histopathology images on four TCGA and two independent cohorts

As illustrated in **Figure 1**, we trained our models with histopathology images and their corresponding normalized gene expression profiles for each of the four TCGA cancer types studied. We then applied the trained models to predict gene expression of the internal held-out test sets. To evaluate model performance, we estimated the Pearson correlation (R) between predicted and actual expression values of each gene across the samples in each test set, taking the mean correlation across all folds. For each cancer type, a total of approximately 18,000 genes were studied; of these, a majority of genes (over 15,000 genes; 80%) were significantly predicted, with Holm-Sidak corrected p-values < 0.05. For each cancer type, over 1,800 genes had a correlation above 0.4 (**Figure 2a-d** and **Extended Table 1**). Benchmarking against two recently published state-of-the-art expression prediction approaches, DeepPT predicted 2,743 genes for lung cancer and 1,812 genes for breast cancer with mean correlations greater than 0.4, roughly doubling the number of genes predicted to this extent (the 'mark' set) by Schmauch et al [29] (1,550 and 786 genes for lung and breast cancer, respectively) (**Extended Table 1**). The scatterplots of the prediction results for the 30 best predicted genes across samples are presented for each dataset (**Extended Figure 3-6**).

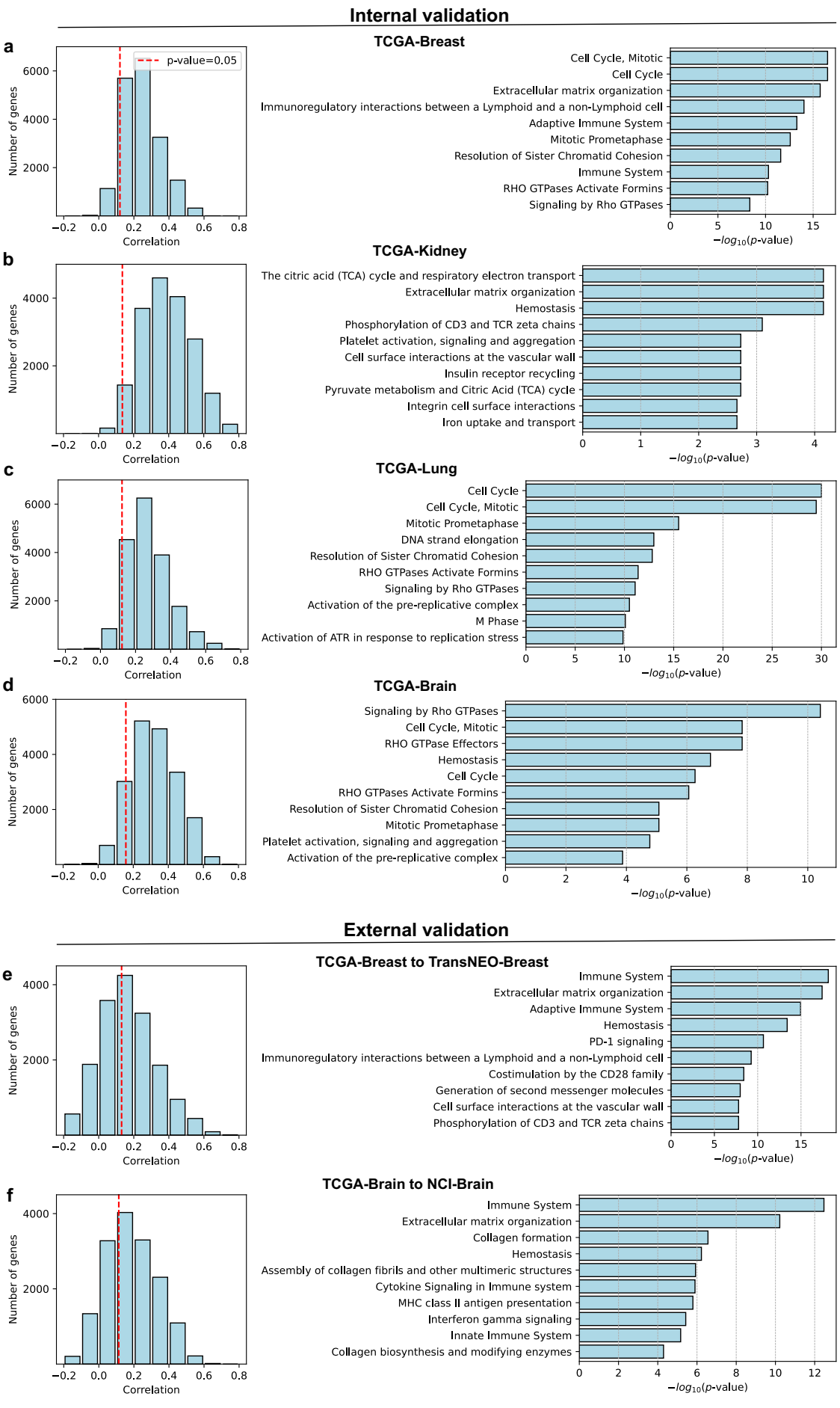

Internal validation
a
TCGA-Breast
p-value=0.05
Number of genes
Correlation
Cell Cycle, Mitotic
Cell Cycle
Extracellular matrix organization
Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell
Adaptive Immune System
Mitotic Prometaphase
Resolution of Sister Chromatid Cohesion
Immune System
RHO GTPases Activate Formins
Signaling by Rho GTPases
−log10(p-value)
b
TCGA-Kidney
The citric acid (TCA) cycle and respiratory electron transport
Extracellular matrix organization
Hemostasis
Phosphorylation of CD3 and TCR zeta chains
Platelet activation, signaling and aggregation
Cell surface interactions at the vascular wall
Insulin receptor recycling
Pyruvate metabolism and Citric Acid (TCA) cycle
Integrin cell surface interactions
Iron uptake and transport
c
TCGA-Lung
Cell Cycle
Cell Cycle, Mitotic
Mitotic Prometaphase
DNA strand elongation
Resolution of Sister Chromatid Cohesion
RHO GTPases Activate Formins
Signaling by Rho GTPases
Activation of the pre-replicative complex
M Phase
Activation of ATR in response to replication stress
d
TCGA-Brain
Signaling by Rho GTPases
Cell Cycle, Mitotic
RHO GTPase Effectors
Hemostasis
Cell Cycle
RHO GTPases Activate Formins
Resolution of Sister Chromatid Cohesion
Mitotic Prometaphase
Platelet activation, signaling and aggregation
Activation of the pre-replicative complex
External validation
e
TCGA-Breast to TransNEO-Breast
Immune System
Extracellular matrix organization
Adaptive Immune System
Hemostasis
PD-1 signaling
Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell
Costimulation by the CD28 family
Generation of second messenger molecules
Cell surface interactions at the vascular wall
Phosphorylation of CD3 and TCR zeta chains
f
TCGA-Brain to NCI-Brain
Immune System
Extracellular matrix organization
Collagen formation
Hemostasis
Assembly of collagen fibrils and other multimeric structures
Cytokine Signaling in Immune system
MHC class II antigen presentation
Interferon gamma signaling
Innate Immune System
Collagen biosynthesis and modifying enzymes

Figure 2. DeepPT prediction of gene expression from H&E slides. Histograms of the Pearson correlation coefficients (R) between predicted and actual expression for each gene across the test set (left panels), and top enriched pathways among the well-predicted genes (right panels). Red dashed lines in the left panels represent the correlation coefficient level beyond which the results are significant (p-value < 0.05 after correction for multiple hypotheses testing).

Notably, a gene set enrichment analysis using the GSEApy python package [38] and Reactome database [39] reveals an enrichment of cellular functions known to play a key role in oncogenesis, such as cell cycle, mitosis, and extracellular matrix organization. Many enriched pathways are specific to different cancer types, pointing to an interesting tumour specificity of the pathways whose expression is associated with H&E features (**Figure 2a-d, right panels**). These results indicate that transcriptomic alterations in oncogenic pathways are indeed more likely to be more strongly associated with visible changes in the H&E slides, but the exact nature of these associations is tumour specific.

As an independent external test, we used the trained models with TCGA-breast cancer to predict gene expression in the TransNEO slides (N=160). Remarkably, the two datasets were generated independently at different facilities, with two different preparation methods (TCGA slides are FFPE, while TransNEO slides are FF). Hence, histological features extracted from these two datasets are clearly distinct (**Extended Figure 7**). Despite these marked data differences, without any further training, we found 1,489 genes with R > 0.4 (**Figure 2e**, **Extended Table 1,** and **Extended Figure 8**). Similarly, we also applied the models trained with TCGA-Brain samples to predict gene expression from NCI-Brain slides (a new unpublished dataset; N=210). We observed 1,324 genes with R > 0.4 (**Figure 2f**, **Extended Table 1,** and **Extended Figure 9**). The strong performance of DeepPT on the external datasets suggests that the model can generalize to new data.

Predicting treatment response from DeepPT-predicted gene expression

In principle, predicting response to treatment directly from H&E slides could be of great value. However, a purely supervised approach, which trains over TCGA data to predict treatment response and then applies the trained model on new data, is not feasible since treatment response data is rare for targeted therapies in TCGA (and more generally such data is scarce and still hard to obtain). Therefore, we instead developed a two-step approach that promises far

wider applicability. First, we applied DeepPT to predict the patient transcriptomics from their H&E slides. Second, based on this predicted gene expression, we used the precision oncology algorithm, ENLIGHT [33], to predict the patients' response from the predicted expression.

Leveraging this computational pipeline, we tested the ability of ENLIGHT to accurately predict patient response in a cohort of 67 HER2+ breast cancer patients who had received a combination of chemotherapy and Trastuzumab (targeting *ERBB2*) as part of the TransNEO trial [37]. Data utilized in this analysis included the response to therapy by residual cancer burden criteria and the fresh frozen H&E-stained primary tumor slides. First, we used the DeepPT model trained beforehand on the 1,043 TCGA-Breast patients, *without any changes* and with no further training on this test dataset, to predict the gene expression values from the H&E slides of each patient's tumor. Second, we applied ENLIGHT to these predicted gene expression values to produce EMS scores based on the same SL/SR partners of *ERBB2* that were already inferred in [33]. Importantly, we do not restrict the model to genes with high correlations between the actual and predicted expression values; ENLIGHT considers the combined effect of a large set of genes, mitigating the noise of individual expression prediction, which leads to robust response predictions as we show below. The only modification made to the original ENLIGHT version is the exclusion of the component that considers the expression of the drug target (*ERBB2*) itself, as this component highly weighs a single gene and is hence much more susceptible to errors in the prediction of expression values.

We compared the prediction accuracy of this approach (termed *ENLIGHT-DeepPT*) to the accuracy of the original ENLIGHT scores when calculated based on actual expression values (*ENLIGHT-actual*), as well as to a multi-omic machine learning predictor that uses DNA, RNA and clinical data, published by Sammut *et al.* in their original study of this dataset (*Sammut-ML*) [37]. **Figure 3a** shows the performance of *ENLIGHT-DeepPT* and *ENLIGHT-actual* predictions in terms of odds ratio (OR, left panel), positive predictive value (PPV; Precision, middle panel) and sensitivity (Recall, right panel). The OR denotes the ratio of the odds to respond among patients predicted to respond vs. the odds to respond among patients predicted not to respond. The PPV denotes the fraction of true responders out of those predicted as such, and the sensitivity denotes the fraction of those predicted to respond out of all true responders. Patients were predicted as *Responders* (or *Non-Responders*) if their EMS scores were greater/equal (lesser) than a decision threshold value of 0.54, a threshold that was fixed and determined already in [33], again, on completely independent data. The same threshold was

used here, without any training or modification, both for *ENLIGHT-DeepPT* and *ENLIGHT-actual*. Using this previously established threshold, the OR of *ENLIGHT-DeepPT* is **5.97**, which is strikingly similar to the OR of **5.96** obtained by *ENLIGHT-actual*. The PPV of *ENLIGHT-DeepPT* was **64%**, higher than the PPV of **52%** when using *ENLIGHT-actual*, and both were much higher than the overall observed response rate of 29.7%. *ENLIGHT-DeepPT* sensitivity was however markedly lower than that of *ENLIGHT-actual*, **37%,** vs. **64%**.

Figure 3b compares ENLIGHT's performance (both versions) to that of the *Sammut-ML* predictor [37]. To recount, the latter was based on *in-cohort supervised learning* to predict response to chemotherapy with or without trastuzumab among HER2+ breast cancer patients. Here, we applied each method to the patients for which the relevant data was available: 65 patients had RNAseq data (*ENLIGHT-actual*), 64 had H&E slides (*ENLIGHT-DeepPT*), and 56 had RNAseq, DNAseq and clinical features needed to run the *Sammut-ML* predictor. First, to systematically compare between the predictors across a wide range of decision thresholds as customary in the literature, we used two overall measures: the area under the ROC curve (AUC), which measures how well a model can rank the samples, and the average precision (AP), which measures how well a model can correctly identify all the true responders without marking too many true non-responders as responders. Second, we aimed to compare these three predictors based on their ability to classify samples as responders or non-responders for practical clinical usage (as we did above when comparing *ENLIGHT-DeepPT* and *ENLIGHT-actual*). Since Sammut et al. did not select a decision threshold, we made this comparison by setting a threshold for each method such that the sensitivity is fixed, once capturing one third of the true responders and once capturing two thirds of them (for confidence intervals and exact values of all measures, see **Extended Table 2**; this further complements the results in **Figure 3a** which use the predefined threshold of 0.54). As can be seen, all methods have quite comparable predictive power, where especially the performance of *ENLIGHT-actual* and *Sammut-ML* predictors is similar across these diverse measures, while that of *ENLIGHT-DeepPT* differs. However, importantly, using only H&E slides which are readily available, without need for RNA or DNA data or other clinical features, has a major practical advantage over the other methods.

The above results support the notion that the DeepPT-based ENLIGHT approach is quite robust to measurement noise, enabling expression predictions on unseen data in an unsupervised manner, without further training. Notably, the predictions were made on fresh

frozen tissue slides, while the DeepPT model was trained on FFPE samples, which differ considerably from FF samples, further testifying to the robustness of DeepPT and ENLIGHT.

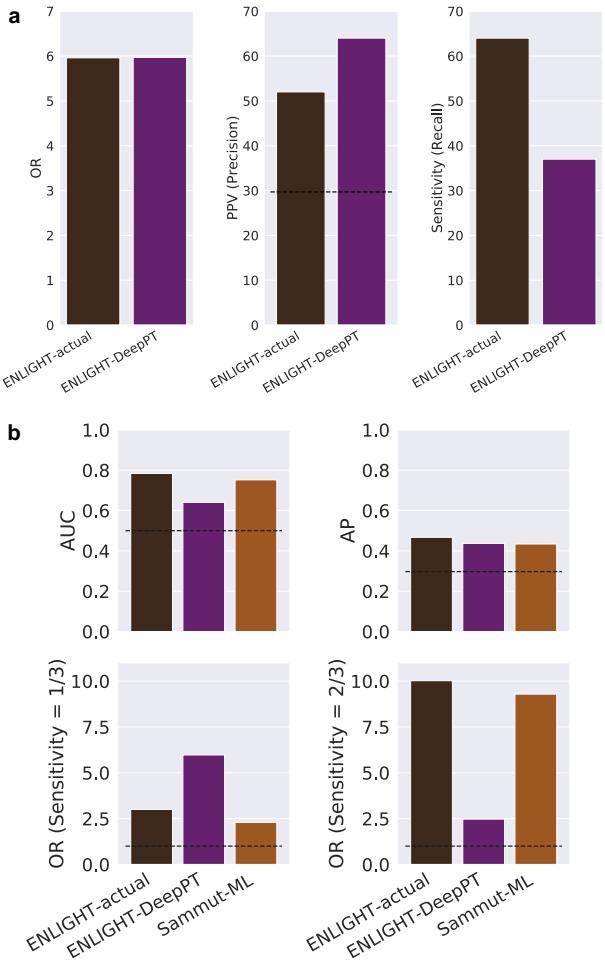


Figure 3. Predicting treatment response from H&E slides. (a) Odds Ratio (OR, left panel), Positive Predictive Value (PPV, middle panel) and Sensitivity (Recall, right panel) of response prediction based on ENLIGHT, using either the actual expression or the DeepPT-predicted expression. Comparison is made on the 64 of the 67 patients for which both RNAseq values and H&E slides were available. The overall response rate (19/64, 29.7%) is denoted by a horizontal dashed line in the center panel. **(b)** Comparison of both ENLIGHT based models and the *Sammut-ML* predictor of Sammut *et al.* [37] using either the actual expression or the DeepPT predicted expression. Each method was applied to the patients for which relevant data is available: 65 patients had RNAseq data, 64 had H&E slides, and 56 had RNAseq and clinical data needed to run the Sammut et al. predictor. Odds Ratio (OR) is calculated at two thresholds for each predictor, one which achieves a sensitivity of one third and one of two thirds. Horizontal dashed lines denote the corresponding baseline value for each measure.

DISCUSSION

Our study demonstrates that an appropriate combination of DeepPT, a novel deep learning framework for predicting gene expression from H&E slides, and ENLIGHT, an unsupervised computational approach for predicting patient response from their predicted tumour gene expression, could be used to form a new approach for H&E-based personalized medicine, in an integrated form termed *ENLIGHT-DeepPT*. We show that *ENLIGHT-DeepPT* successfully predicts the true responders in a recently published multi-omic breast cancer clinical trial directly from the H&E images, obtaining a clinically meaningful hazard ratio of about six.

DeepPT is fundamentally different from previous computational pipelines for gene expression prediction both in its model architecture and in its training strategy. We attribute its superior performance to four key innovations: (i) All previous studies fed output from the conventional pre-trained CNN model (with natural images from ImageNet database) directly into their regression module, whereas we added an auto-encoder to re-train the output of the pre-trained CNN model. This helps to familiarise the model with histological features, to exclude noise, to avoid over fitting, and finally to reduce the computational demands. (ii) Our regression module is a non-linear MLP model in which the weights from the input layer to the hidden layer are shared among genes. This architecture enables the model to exploit the correlations between the expression of the genes. (iii) We trained together sets of genes with similar median gene expression values; doing so further implements a form of multitask learning and prevents the model from focusing on only the most highly expressed genes. (iv) We performed ensemble learning by taking the mean predictions across all models. This further improves the prediction accuracy quite significantly (**Extended Figure 2c**).

DeepPT can be broadly applied to other cancer types; however, similar to many other deep learning models, it requires a sufficient number of training samples. An interesting direction for future work would be to apply transfer learning between cohorts, to improve the predictive performance in cancer types with small training cohorts. In other words, it might be possible to train the model on large datasets such as breast and lung cancer, then fine tune it for generating predictions in smaller datasets such as pancreatic cancer or melanoma.

ENLIGHT is an unsupervised approach that leverages large-scale data in cancer to infer genetic interaction partners associated with drug targets, and then uses their activation patterns

in the tumor to generate a matching score for each possible treatment, given a tumor sample. The ENLIGHT matching score is translated to a binary response prediction using a single pan-cancer pan-treatment threshold, above which a treatment is considered a match to the patient. The ENLIGHT framework was originally developed and validated using RNA expression data, and was applied here *as-is* to DeepPT-predicted expression, generating *ENLIGHT-DeepPT*. The robustness of the overall approach was demonstrated in an independent test set, where the response prediction accuracy was on par with that achieved using the actual expression data, as well as that of a supervised classifier trained for this task that was published in the original study. Combining DeepPT with ENLIGHT is a promising approach to predicting response to treatment directly from H&E slides because it does not require response data on which to train. This is a crucial advantage compared to the more common practice of using response data to train classifiers in a supervised manner. Indeed, while the TCGA lacks response data that would enable a supervised predictor of response to Trastuzumab, applying ENLIGHT on predicted expression here has successfully enabled the prediction of response to Trastuzumab in the TransNEO dataset with considerable accuracy. An additional motivation for developing a response prediction pipeline directly from H&E slides is that NGS results often take 4-6 weeks after initiation to return a result. Many patients who have advanced cancers require treatment immediately, and this method can potentially offer treatment options within a short time frame.

A striking finding of this study is the robustness of response predictions based on H&E slides when combining DeepPT and ENLIGHT. First, despite the inevitable noise introduced by the prediction of gene expression, the original ENLIGHT inferred GI partners of Trastuzumab did not require any modifications to predict response here from the DeepPT-predicted expression. Second, though DeepPT was trained using FFPE slides, it generalized well and could be used as-is to predict expression values from FF slides. This demonstrates the applicability of DeepPT for predicting RNA expression either from FF or from FFPE slides. Nevertheless, a limitation of this work is that additional validation will be required, once appropriate publicly available resources become available. We are currently focusing on obtaining and analysing additional test cohorts for an extended journal submission of this manuscript, but we publish the current interim results on bioRxiv, given their potential interest.

The growing efforts to harness the rapid advances in deep learning to improve cancer patients care are obviously laudable. Those include a variety of promising studies developing

new methods to classify tumors, predict their survival and their response to therapy from tumor slides. The vast majority of current studies aimed at predicting patient response have been focused on predicting genomic alterations; to date, we are not aware of studies aiming at predicting patient response to therapy from the predicted tumor expression from histological images. Given its general and unsupervised nature, we are hopeful that *ENLIGHT-DeepPT* may possibly have considerable impact, making precision oncology more accessible to patients in the developing world and in other situations where sequencing is less feasible. While promising, one should cautiously note that the results presented await a broader validation.

METHODS

Data collection

The datasets in this study were collected from three resources: TCGA, TransNEO, and Laboratory of Pathology at the NCI.

The TCGA histological images and their corresponding gene expression profiles were downloaded from the TCGA database (https://portal.gdc.cancer.gov). Only diagnostic slides from primary tumor were selected, making a total of 4,368 formalin-fixed paraffin-embedded (FFPE) slides from 3,750 patients with breast cancer (1,106 slides; 1,043 patients), kidney cancer (859 slides; 836 patients), lung cancer (1,018 slides; 927 patients), and brain cancer (1,015 slides; 574 patients).

The TransNEO-Breast dataset consists of fresh frozen (FF) slides and their corresponding gene expression profiles from 160 breast cancer patients. Full details of the RNA library preparation and sequencing protocols, as well as digitisation of slides have been previously described [37].

The NCI-Brain histological images and their corresponding gene expression profiles were obtained from archives of the Laboratory of Pathology at the NCI, and consisted of 210 cases comprising a variety of CNS tumors, including both common and rare tumor entities. All cases were subject to methylation profiling to evaluate the diagnosis, as well as RNA-sequencing.

Histopathology image processing

We first used method of Otsu [40] to identify areas containing tissue within each slide. Because the WSI are too large (from 10,000 to 100,00 pixels in each dimension) to feed directly into the deep neural networks, we then partitioned the WSI at 20x magnification into non-overlapping tiles comprised of 512 x 512 RGB pixels. Tiles containing heavy marks or more than 50% of background were removed. Depending on the size of slide, the number of tiles per slide in TCGA cohort varied from 100 to 8,000 (**Extended Figure 10a-d**). In contrast, TransNEO slides are much smaller, resulting in 100 to 1,000 tiles per slide (**Extended Figure 10e**). To minimize staining variation (heterogeneity and batch effects), color normalization was applied for the selected tiles [41–43].

Gene expression processing

Gene expression profiles were obtained from read counts for approximately 60,000 gene identifiers. A subset of highly expressed genes was identified using edgeR, resulting in roughly 18,000 genes for each cancer type. The median expression across samples of each gene varied from 10 to 10,000 across for every dataset (**Extended Figure 11**). To reduce the range of gene expression values, and to minimize discrepancies in library size between experiments and batches, a normalization was performed as described in our previous work [33].

Model architecture

Our model architecture was composed of three main units (**Extended Fig. 1**).

(1) Feature extraction: The pre-trained ResNet50 CNN model with 14 million natural images from the ImageNet database [34] was used to extract features from image tiles. Before feeding these tiles into the ResNet50 unit, the image tiles were resized to 224 x 224 pixels to match the standard input size for the convolutional neural network. Through the feature extraction process, each input tile is represented by a vector of 2,048 derived features.

(2) Feature compression: We applied an autoencoder, which consists of a bottleneck of 512 neurons, to reduce the number of features from 2,048 to 512. This helps to familiarise the model with histological features, to exclude noise, to avoid over fitting, and finally to reduce the computational demands.

(3) Multi-Layer Perceptron (MLP) regression: The purpose of this component is to build a predictive model linking the aforementioned auto-encoded features to whole-genome gene expression. The model consists of three layers: (1) an input layer with 512 nodes, reflecting the size of the auto-encoded vector; (2) a hidden layer whose size depends on the number of genes under shared consideration; and (3) an output layer with one node per gene. The rationale behind this architecture is to leverage similarity among the genes under shared consideration, as captured by the weights connecting the input layer to the hidden layer. The weights connecting the hidden layer to the output layer model the subsequent relationship between the hidden layer and each individual gene. This follows the philosophy of multi-task learning. If the prediction of each gene's expression level represents a single task, then our strategy is to first group these tasks for shared learning, followed by optimization of each individual task. In

our default whole-genome approach, we bin genes into groups of 4,096 whose median expression levels are similar, and we use 512 hidden nodes. Because the training data is comprised of gene expression at the slide level (i.e. bulk gene expression, as opposed to at spatial resolution), we average our per-tile predictions to obtain bulk values at the slide level.

Model training and evaluation

We trained and evaluated each cancer type independently. To evaluate our model performance, we applied 5x5 nested cross-validation. For each outer loop, we split the entire samples (of each cohort) into training (80%) and held-out test (20%) set. We further split the training set into internal training and evaluation set, according five-fold cross validation. The models were trained and evaluated independently with the different pairs of training/validation sets. Averaging the predictions from the five different models represents our final prediction for each single gene on each held-out test set. We repeated this procedure five times across the five held-out test sets, making a total of 25 trained models. These models trained with TCGA cohorts were used to predict the expression of each gene in a given external cohort by computing the mean over the predicted values of all models.

As noted in the Model Architecture section, tranches of genes with similar median expression levels were grouped for simultaneous training and evaluation. This was done to optimize model performance and model efficiency, and contrasts approaches in the literature that either train on each gene separately [27] or on all genes together [29]. Each training round was stopped at a maximum of 500 epochs, or sooner if the average correlation per gene between actual and prediction values of gene expression on the validation set did not improve for 50 continuous epochs. The Adam optimizer with mean squared error loss function was employed in both auto-encoder and MLP models. A learning rate of 10^{-4} and a minibatches of 32 image tiles per step were used for both the auto-encoder model and MLP regression model. To avoid overfitting, a dropout of 0.2 was also used.

Implementation details

All analysis in this study was performed in Python 3.7.4 and R 4.1.0 with the libraries including Numpy 1.18.5, Pandas 1.0.5, Scikit-learn 0.23.1, Matplotlib 3.2.2, and edgeR 3.28.0. Image processing including tile partitioning and color normalization was conducted with OpenSlide 1.1.2, OpenCV 4.4.0, PIL 6.1.0, and colorcorrect 0.9.1. The histopathology feature extraction

was carried out using TensorFlow 2.2.0. The feature compression (autoencoder unit) and MLP regression parts were implemented using PyTorch 1.7.0. Pearson correlation was calculated using Scipy 1.5.0. Gene set enrichment analysis was performed with Gseapy 0.10.7. The identification of highly expressed genes was performed with edgeR 3.28.0.

Acknowledgments

Many thanks to Drs. Stephen-John Sammut and Carlos Caldas for generously making their clinical trial data available to us and for very helpful discussions and advice. This work was supported by the Australian Research Council (ARC) (D.T.H, E.A.S) and by the Intramural Research Program of the National Institutes of Health (NIH), National Cancer Institute (NCI), Center for Cancer Research (CCR) (S.S, N.R, E.R). This work utilized the super computational resources of the Australian National Computational Infrastructure (AUNCI) and the Australian National University Merit Allocation Scheme (ANUMAS).

Declaration of interests

G.D, D.S.B, E.E, T.B, and R.S are employees of Pangea Biomed. E.R. is a co-founder of Pangea Biomed.

References

1. Ström P, Kartasalo K, Olsson H, Solorzano L, Delahunt B, Berney DM, et al. Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study. Lancet Oncol. 2020;21: 222–232.

2. Fu Y, Jung AW, Torne RV, Gonzalez S, Vöhringer H, Shmatko A, et al. Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. Nat Cancer. 2020;1: 800–810.

3. Noorbakhsh J, Farahmand S, Foroughi Pour A, Namburi S, Caruana D, Rimm D, et al. Deep learning-based cross-classifications reveal conserved spatial behaviors within tumor histological images. Nat Commun. 2020;11: 6367.

4. Coudray N, Ocampo PS, Sakellaropoulos T, Narula N, Snuderl M, Fenyö D, et al. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. Nat Med. 2018;24: 1559–1567.

5. Swiderska-Chadaj Z, Pinckaers H, van Rijthoven M, Balkenhol M, Melnikova M, Geessink O, et al. Learning to detect lymphocytes in immunohistochemistry with deep learning. Medical Image Analysis. 2019. p. 101547. doi:10.1016/j.media.2019.101547

6. Yu K-H, Wang F, Berry GJ, Ré C, Altman RB, Snyder M, et al. Classifying non-small cell lung cancer types and transcriptomic subtypes using convolutional neural networks. J Am Med Inform Assoc. 2020;27: 757–769.

7. Couture HD, Williams LA, Geradts J, Nyante SJ, Butler EN, Marron JS, et al. Image analysis with deep learning to predict breast cancer grade, ER status, histologic subtype, and intrinsic subtype. NPJ Breast Cancer. 2018;4: 30.

8. Campanella G, Hanna MG, Geneslaw L, Miraflor A, Werneck Krauss Silva V, Busam KJ, et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat Med. 2019;25: 1301–1309.

9. Xu H, Park S, Clemenceau JR, Choi J, Radakovich N, Lee SH, et al. Spatial heterogeneity and organization of tumor mutation burden and immune infiltrates within tumors based on whole slide images correlated with patient survival in bladder cancer. doi:10.1101/554527

10. Kather JN, Heij LR, Grabsch HI, Loeffler C, Echle A, Muti HS, et al. Pan-cancer image-based detection of clinically actionable genetic alterations. Nat Cancer. 2020;1: 789–799.

11. Qu H, Zhou M, Yan Z, Wang H, Rustgi VK, Zhang S, et al. Genetic mutation and biological pathway prediction based on whole slide images in breast carcinoma using deep learning. NPJ Precis Oncol. 2021;5: 87.

12. Schaumberg AJ, Rubin MA, Fuchs TJ. H&E-stained Whole Slide Image Deep Learning Predicts SPOP Mutation State in Prostate Cancer. doi:10.1101/064279

13. Tsou P, Wu C-J. Mapping Driver Mutations to Histopathological Subtypes in Papillary Thyroid Carcinoma: Applying a Deep Convolutional Neural Network. J Clin Med Res. 2019;8. doi:10.3390/jcm8101675

14. Chang P, Grinband J, Weinberg BD, Bardis M, Khy M, Cadena G, et al. Deep-Learning Convolutional Neural Networks Accurately Classify Genetic Mutations in Gliomas. AJNR Am J Neuroradiol. 2018;39: 1201–1207.

15. Kather JN, Pearson AT, Halama N, Jäger D, Krause J, Loosen SH, et al. Deep learning can

predict microsatellite instability directly from histology in gastrointestinal cancer. Nat Med. 2019;25: 1054–1056.

16. Kim RH, Nomikou S, Coudray N, Jour G, Dawood Z, Hong R, et al. A Deep Learning Approach for Rapid Mutational Screening in Melanoma. doi:10.1101/610311

17. Chen M, Zhang B, Topatana W, Cao J, Zhu H, Juengpanich S, et al. Classification and mutation prediction based on histopathology H&E images in liver cancer using deep learning. npj Precision Oncology. 2020. doi:10.1038/s41698-020-0120-3

18. Mobadersany P, Yousefi S, Amgad M, Gutman DA, Barnholtz-Sloan JS, Velázquez Vega JE, et al. Predicting cancer outcomes from histology and genomics using convolutional networks. Proc Natl Acad Sci U S A. 2018;115: E2970–E2979.

19. Cheng J, Zhang J, Han Y, Wang X, Ye X, Meng Y, et al. Integrative Analysis of Histopathological Images and Genomic Data Predicts Clear Cell Renal Cell Carcinoma Prognosis. Cancer Res. 2017;77: e91–e100.

20. Beck AH, Sangoi AR, Leung S, Marinelli RJ, Nielsen TO, van de Vijver MJ, et al. Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. Sci Transl Med. 2011;3: 108ra113.

21. Bulten W, Pinckaers H, van Boven H, Vink R, de Bel T, van Ginneken B, et al. Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study. Lancet Oncol. 2020;21: 233–241.

22. Courtiol P, Maussion C, Moarii M, Pronier E, Pilcer S, Sefta M, et al. Deep learning-based classification of mesothelioma improves prediction of patient outcome. Nat Med. 2019;25: 1519–1525.

23. Zheng H, Momeni A, Cedoz P-L, Vogel H, Gevaert O. Whole slide images reflect DNA methylation patterns of human tumors. NPJ Genom Med. 2020;5: 11.

24. Wang H, Cruz-Roa A, Basavanhally A, Gilmore H, Shih N, Feldman M, et al. Mitosis detection in breast cancer pathology images by combining handcrafted and convolutional neural network features. J Med Imaging (Bellingham). 2014;1: 034003.

25. Turkki R, Linder N, Kovanen PE, Pellinen T, Lundin J. Antibody-supervised deep learning for quantification of tumor-infiltrating immune cells in hematoxylin and eosin stained breast cancer samples. Journal of Pathology Informatics. 2016. p. 38. doi:10.4103/2153-3539.189703

26. He B, Bergenstråhle L, Stenbeck L, Abid A, Andersson A, Borg Å, et al. Integrating spatial gene expression and breast tumour morphology via deep learning. Nat Biomed Eng. 2020;4: 827–834.

27. Wang Y, Kartasalo K, Weitz P, Ács B, Valkonen M, Larsson C, et al. Predicting Molecular Phenotypes from Histopathology Images: A Transcriptome-Wide Expression-Morphology Analysis in Breast Cancer. Cancer Res. 2021;81: 5115–5126.

28. Monjo T, Koido M, Nagasawa S, Suzuki Y, Kamatani Y. Efficient prediction of a spatial transcriptomics profile better characterizes breast cancer tissue sections without costly experimentation. Sci Rep. 2022;12: 4133.

29. Schmauch B, Romagnoni A, Pronier E, Saillard C, Maillé P, Calderaro J, et al. A deep learning model to predict RNA-Seq expression of tumours from whole slide images. Nat Commun. 2020;11: 3877.

30. Levy-Jurgenson A, Tekpli X, Kristensen VN, Yakhini Z. Spatial transcriptomics inferred from pathology whole-slide images links tumor heterogeneity to survival in breast and lung cancer. Sci

Rep. 2020;10: 18802.

31. Echle A, Rindtorff NT, Brinker TJ, Luedde T, Pearson AT, Kather JN. Deep learning in cancer pathology: a new generation of clinical biomarkers. British Journal of Cancer. 2021. pp. 686–696. doi:10.1038/s41416-020-01122-x

32. De Smet F, Antoranz Martinez A, Bosisio FM. Next-Generation Pathology by Multiplexed Immunohistochemistry. Trends Biochem Sci. 2021;46: 80–82.

33. Dinstag G, Shulman ED, Elis E, Ben-Zvi DS, Tirosh O, Maimon E, et al. Clinically oriented prediction of patient response to targeted and immunotherapies from the tumor transcriptome. 2022. doi:10.1101/2022.02.27.481627

34. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016. doi:10.1109/cvpr.2016.90

35. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. ImageNet: A large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition. 2009. doi:10.1109/cvpr.2009.5206848

36. Lee JS, Nair NU, Dinstag G, Chapman L, Chung Y, Wang K, et al. Synthetic lethality-mediated precision oncology via the tumor transcriptome. Cell. 2021;184: 2487–2502.e13.

37. Sammut S-J, Crispin-Ortuzar M, Chin S-F, Provenzano E, Bardwell HA, Ma W, et al. Multi-omic machine learning predictor of breast cancer therapy response. Nature. 2022;601: 623–629.

38. gseapy. In: PyPI [Internet]. [cited 27 May 2022]. Available: https://pypi.org/project/gseapy/

39. Gillespie M, Jassal B, Stephan R, Milacic M, Rothfels K, Senff-Ribeiro A, et al. The reactome pathway knowledgebase 2022. Nucleic Acids Res. 2022;50: D687–D692.

40. Otsu N. A Threshold Selection Method from Gray-Level Histograms. IEEE Transactions on Systems, Man, and Cybernetics. 1979. pp. 62–66. doi:10.1109/tsmc.1979.4310076

41. Rizzi A, Gatta C, Marini D. A new algorithm for unsupervised global and local color correction. Pattern Recognition Letters. 2003. pp. 1663–1677. doi:10.1016/s0167-8655(02)00323-9

42. Lam H-K, Au OC, Wong C-W. Automatic white balancing using luminance component and standard deviation of RGB components [image preprocessing]. 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing. doi:10.1109/icassp.2004.1326589

43. Nikitenko D, Wirth M, Trudel K. Applicability of White-Balancing Algorithms to Restoring Faded Colour Slides: An Empirical Evaluation. Journal of Multimedia. 2008. doi:10.4304/jmm.3.5.9-18

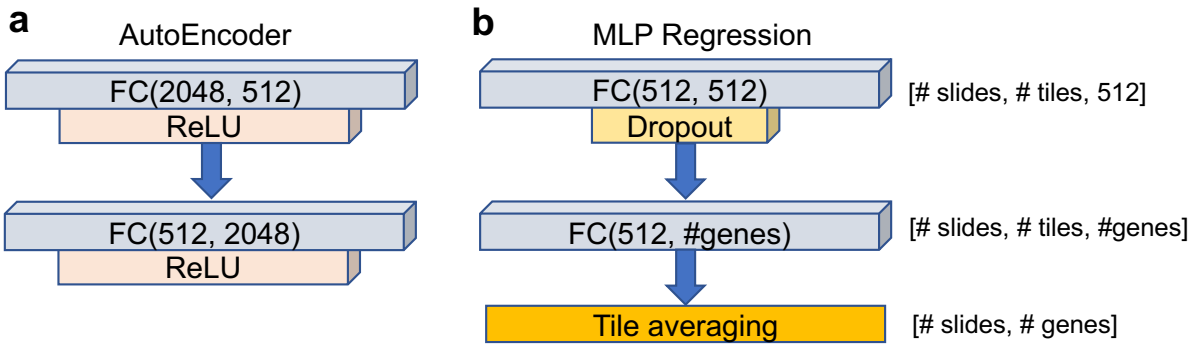
SUPPLEMENTARY DATA

Cohort	Number of patients	Number of genes	
		P-value < 0.05	**R > 0.4**
TCGA-Breast	1,043	15,124 (82%)	1,812
TCGA-Kidney	836	17,326 (95%)	8,312
TCGA-Lung	927	15,361 (84%)	2,743
TCGA-Brain	574	15,531 (81%)	5,341
TransNEO-Breast	160	9,548 (56%)	1,489
NCI-Brain	210	10,386 (66%)	1,324

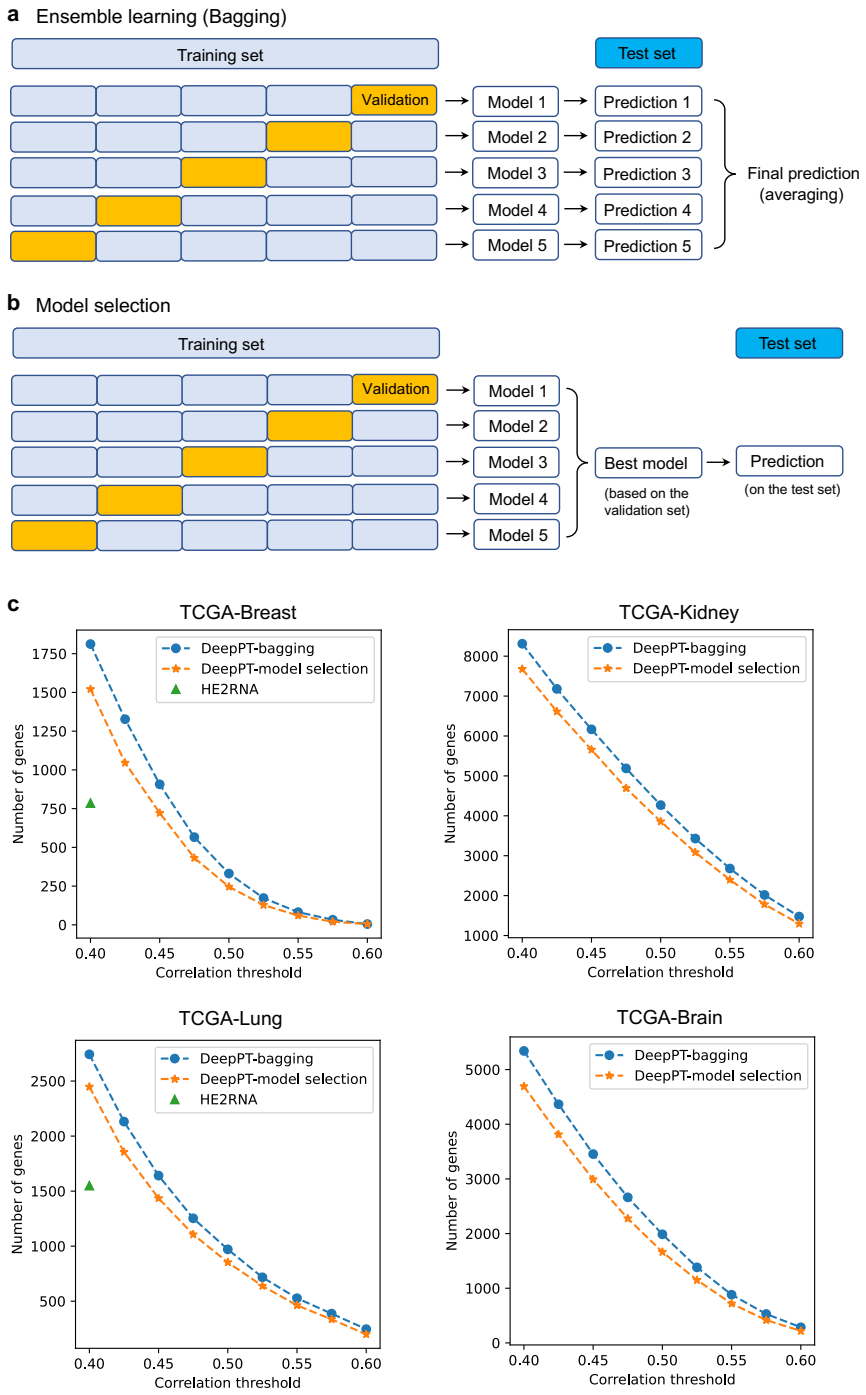
Extended Table 1 | Number of well-predicted genes for each cohort. P-values were adjusted for multiple hypotheses testing using the Holm-Sidak method.

	Number of patients scored (out of 67)	**AUC**	**AP**	**OR (Sensitivity =⅓)**	**OR (Sensitivity =⅔)**
ENLIGHT on actual expression (ENLIGHT-actual)	**65**	**0.784**	**0.467**	3 [0.82,10.94]	**10.02** [2.92,34.37]
ENLIGHT on DeepPT predicted expression (ENLIGHT-DeepPT*)*	64	0.64	0.434	**5.97** [1.49,23.92]	2.47 [0.79,7.67]
Sammut *et al.* on actual expression (*Sammut-ML*)	56	0.752	0.437	2.29 [0.58,8.91]	9.29 [2.53,34.15]

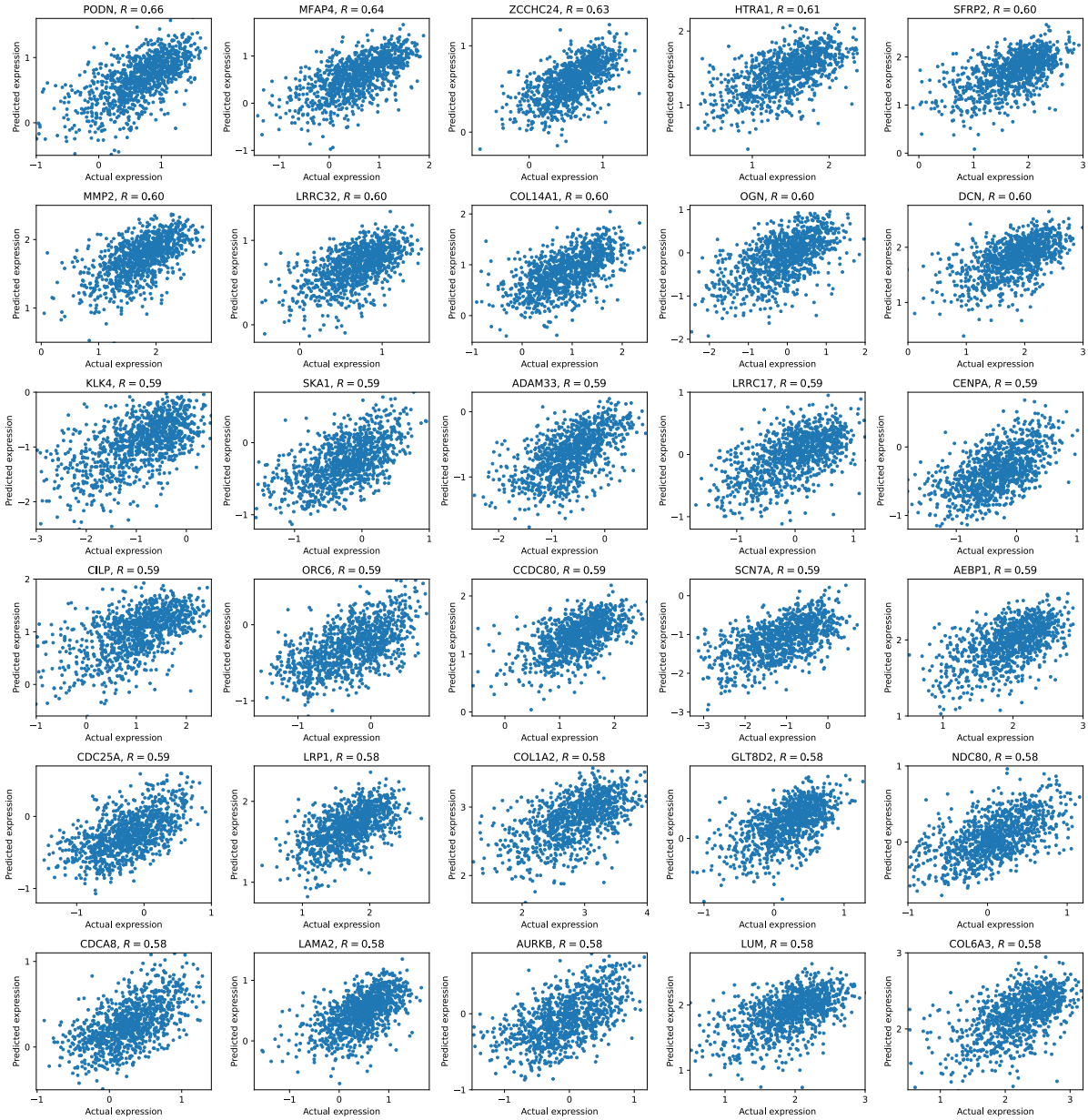
Extended Table 2 | Comparison between methods for predicting treatment response with various methods. AUC: Area under the ROC curve. AP: Average Precision. OR: odds ratio. Brackets indicate the 95% confidence interval. See main text for more details.



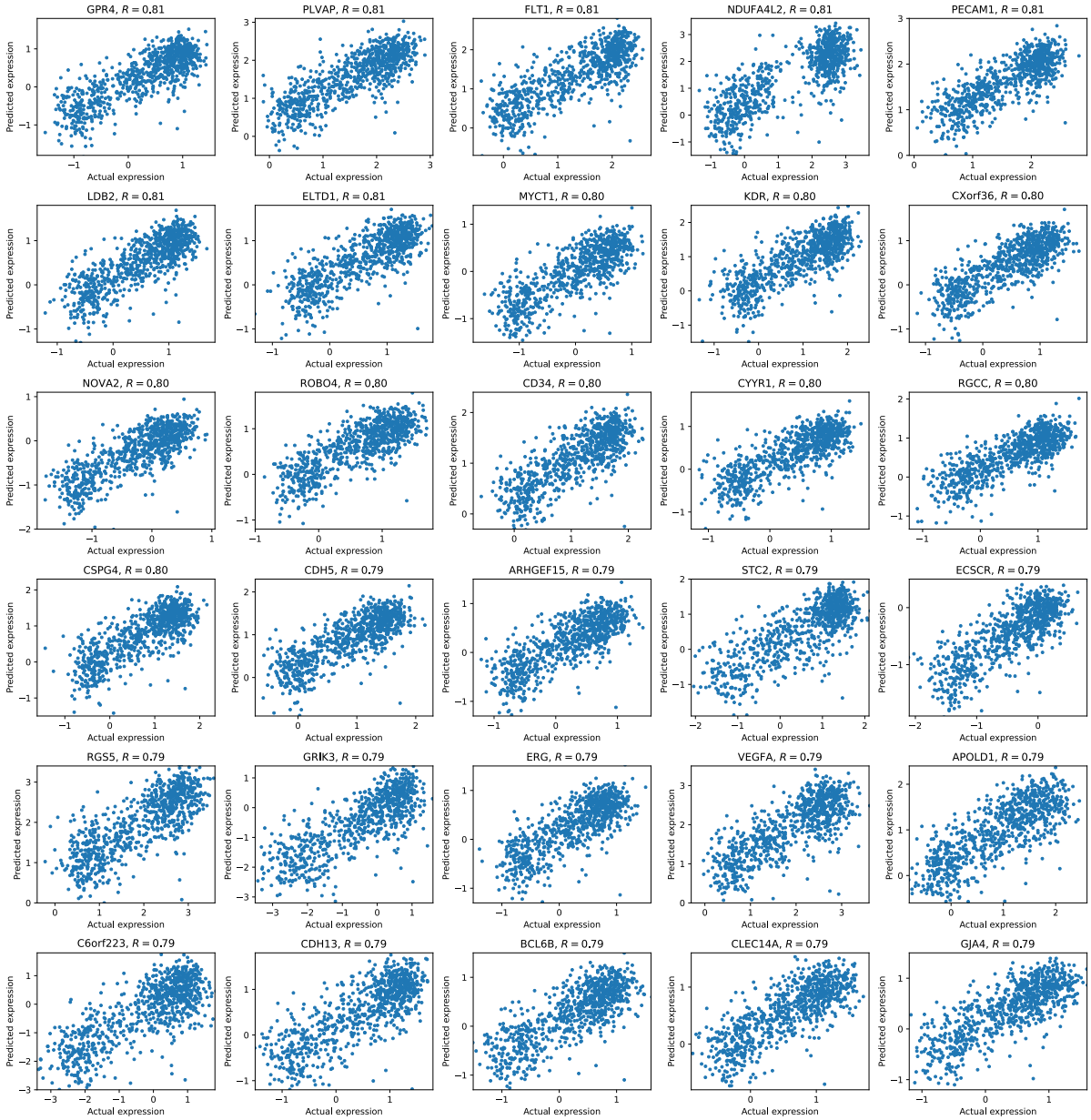
Extended Figure 1 | Model architecture in detail. (a) The feature compression subnetwork consists of an input layer of 2,048 neurons, a bottleneck of 512 neurons, and an output layer of 2,048 neurons. (b) The MLP regression subnetwork consists of an input layer of 512 neurons, a hidden layer of 512 neurons, and an output layers with the number of neurons reflecting the number of genes in each group.



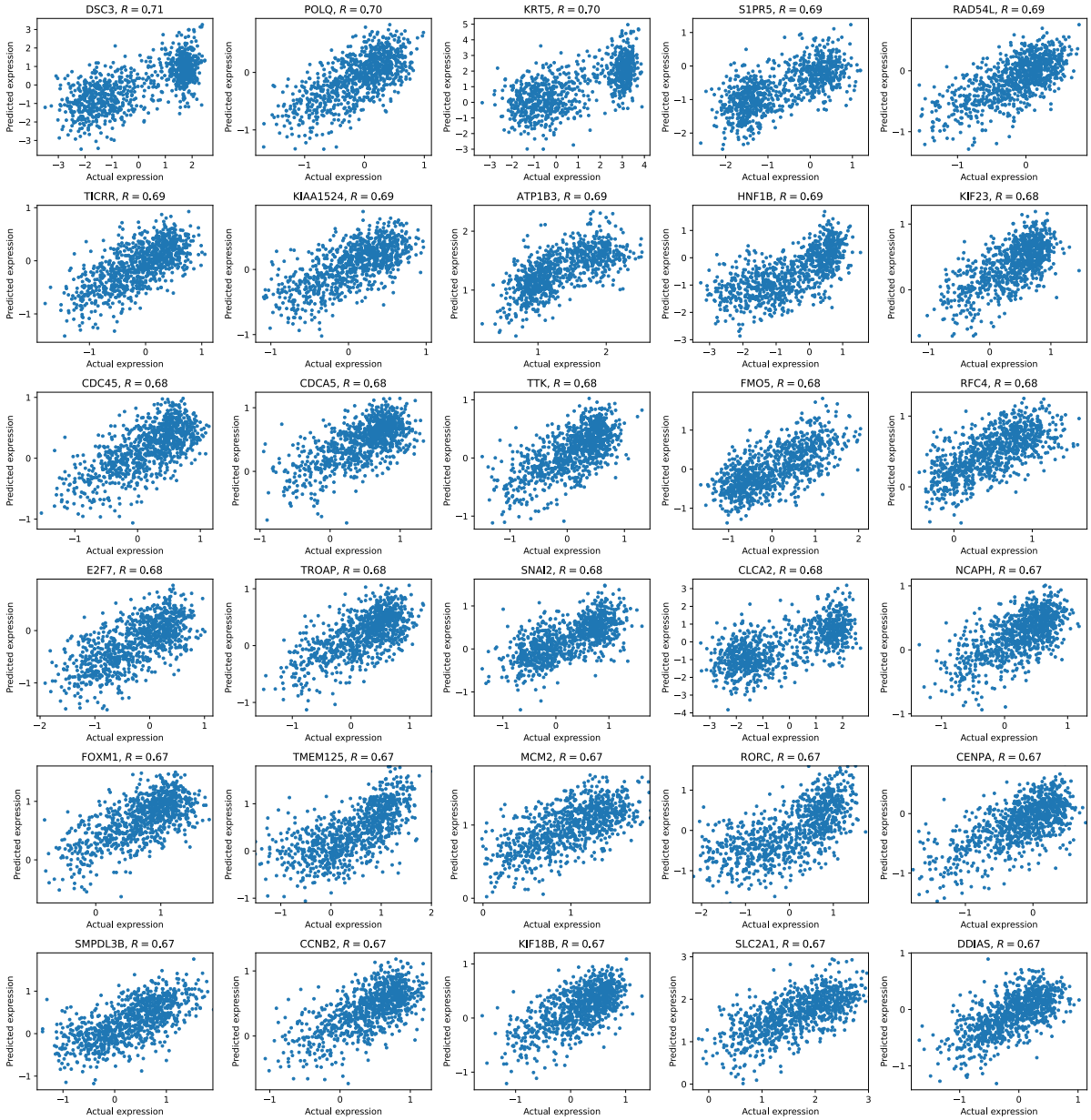
Extended Figure 2 | Training strategies and their performance. In the ensemble learning strategy (bagging), five models were trained independently with five internal training-validation splits; these five model predictions were averaged to make the final prediction (a). In the model selection strategy, the "best" model with the highest performance on the validation set was chosen to make prediction on the test set (b). The comparative performance of these strategies is shown for each cohort (c). Note that with either strategy, DeepPT outperforms the current state-of-the-art approach, HE2RNA.



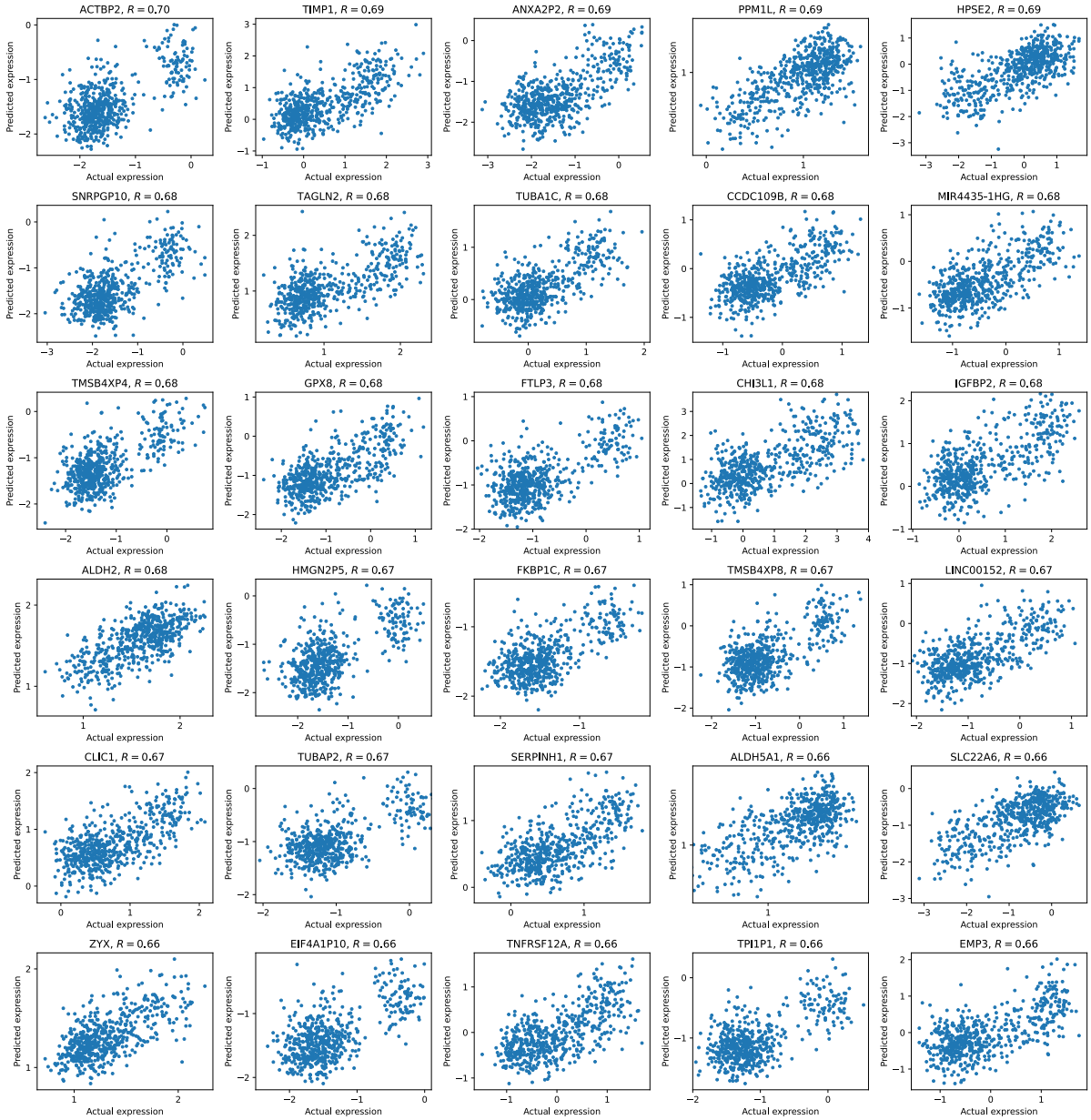
Extended Figure 3 | DeepPT performance on prediction of gene expression for the best thirty genes in TCGA-Breast cohort as measured by Pearson correlation (R). Each scatter plot shows predicted versus actual expression for a single gene across all 1,043 patients.



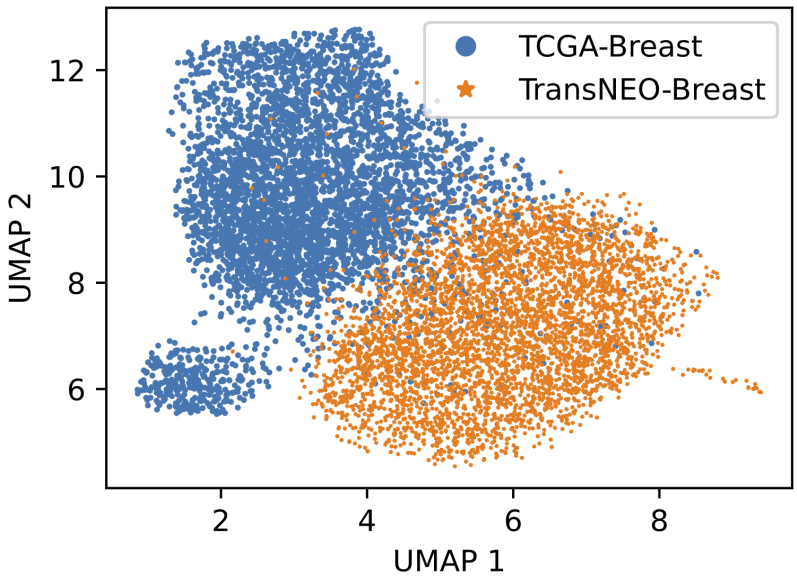
Extended Figure 4 | DeepPT performance on prediction of gene expression for the best thirty genes in TCGA-Kidney cohort as measured by Pearson correlation (R). Each scatter plot shows predicted versus actual expression for a single gene across all 836 patients.



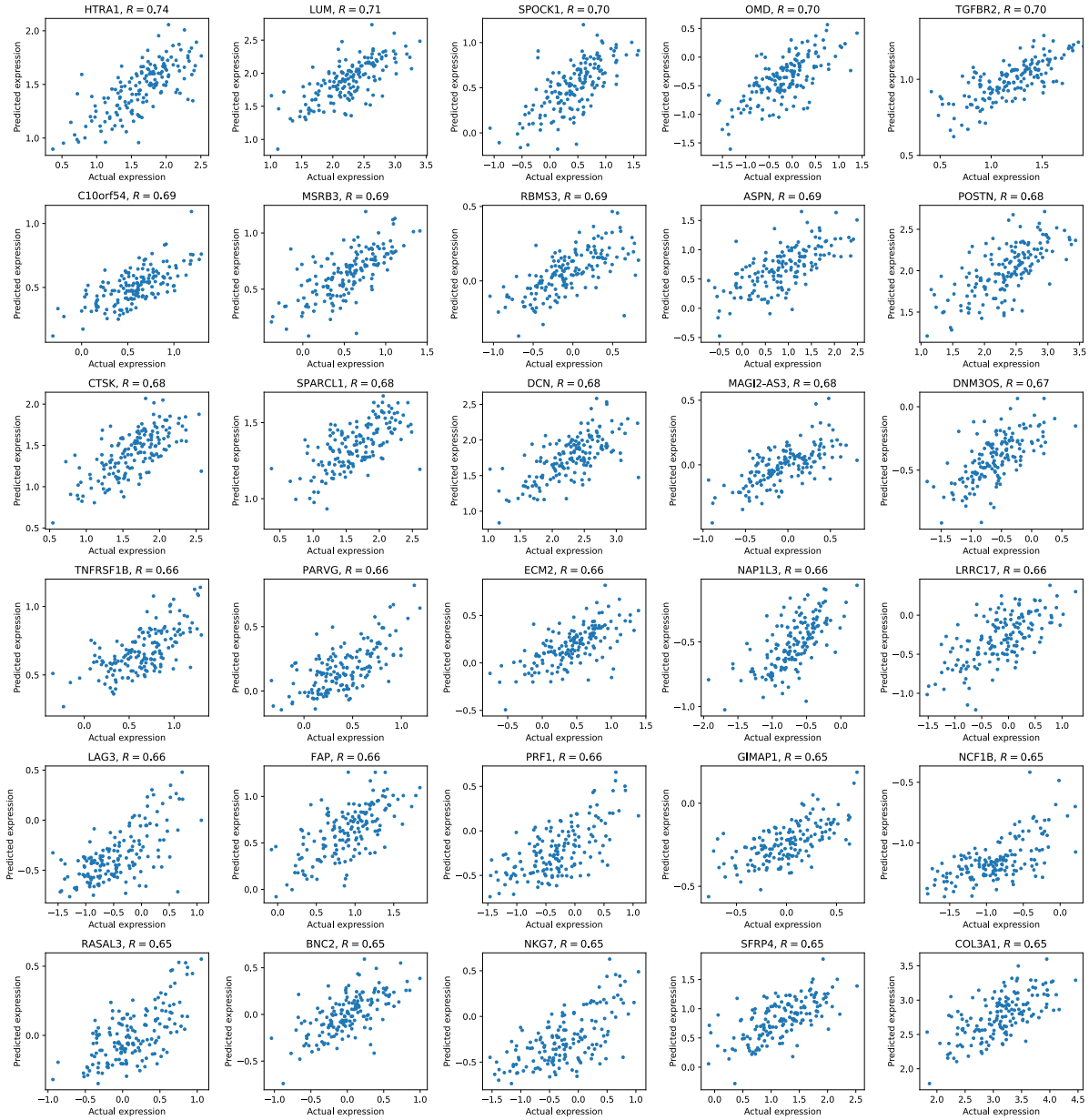
Extended Figure 5 | DeepPT performance on prediction of gene expression for the best thirty genes in TCGA-Lung cohort as measured by Pearson correlation (R). Each scatter plot shows predicted versus actual expression for a single gene across all 927 patients.



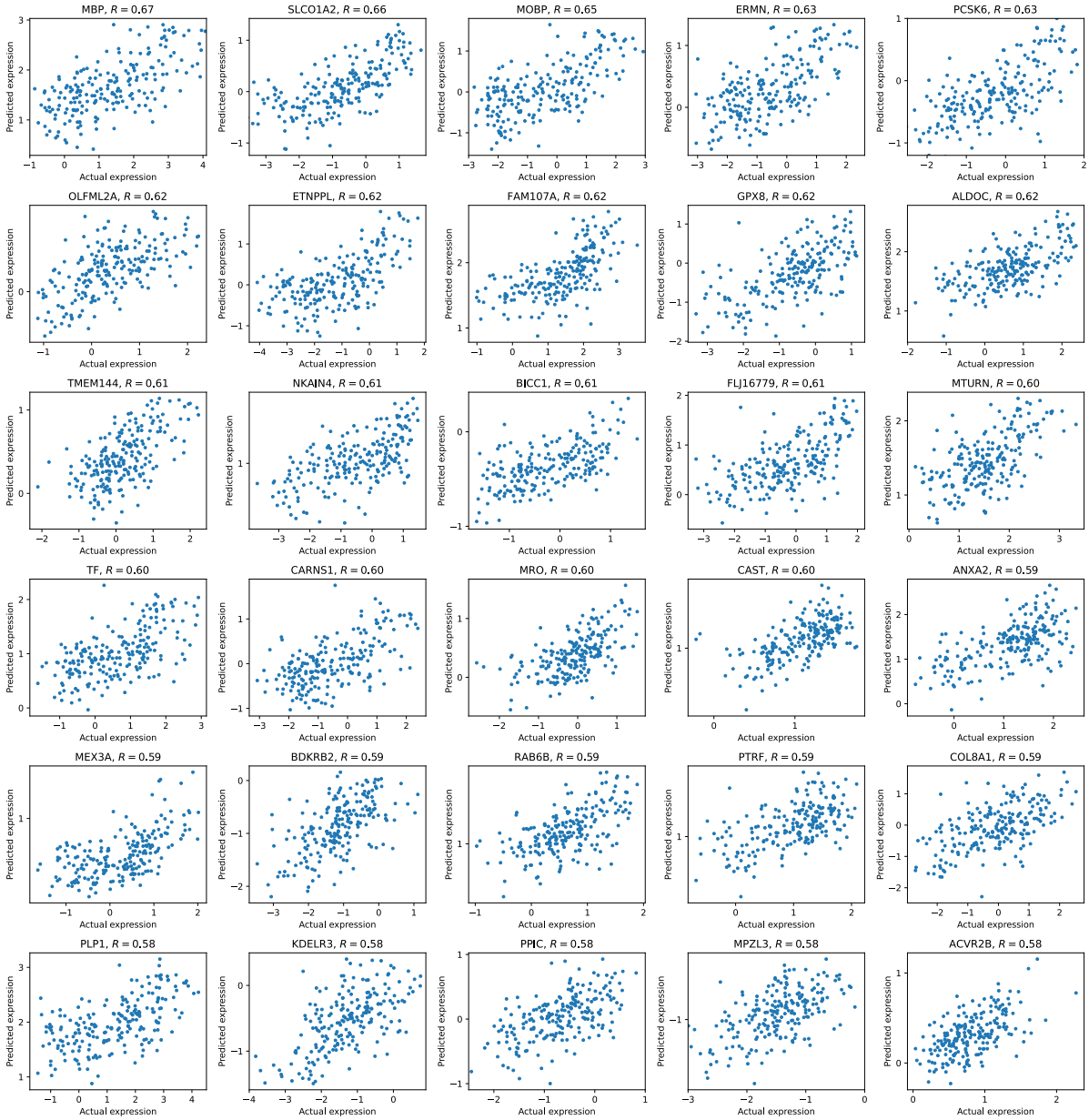
Extended Figure 6 | DeepPT performance on prediction of gene expression for the best thirty genes in TCGA-Brain cohort as measured by Pearson correlation (R). Each scatter plot shows predicted versus actual expression for a single gene across all 574 patients.



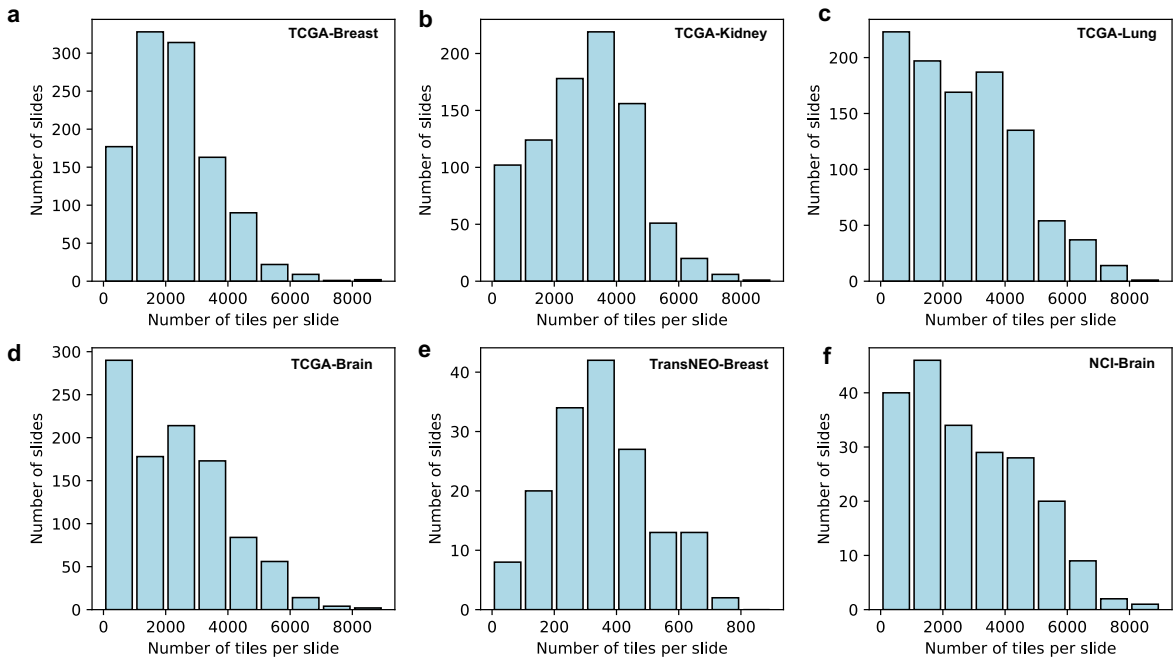
Extended Figure 7 | Difference between histopathological features extracted from TCGA-Breast tiles and TransNEO-Breast tiles. UMAP visualization of 2,048 histopathological features that were extracted by using pre-trained ResNet50 CNN. 4,000 image tiles from each dataset were selected randomly to illustrate. Each point represents each feature vector of one image tile.



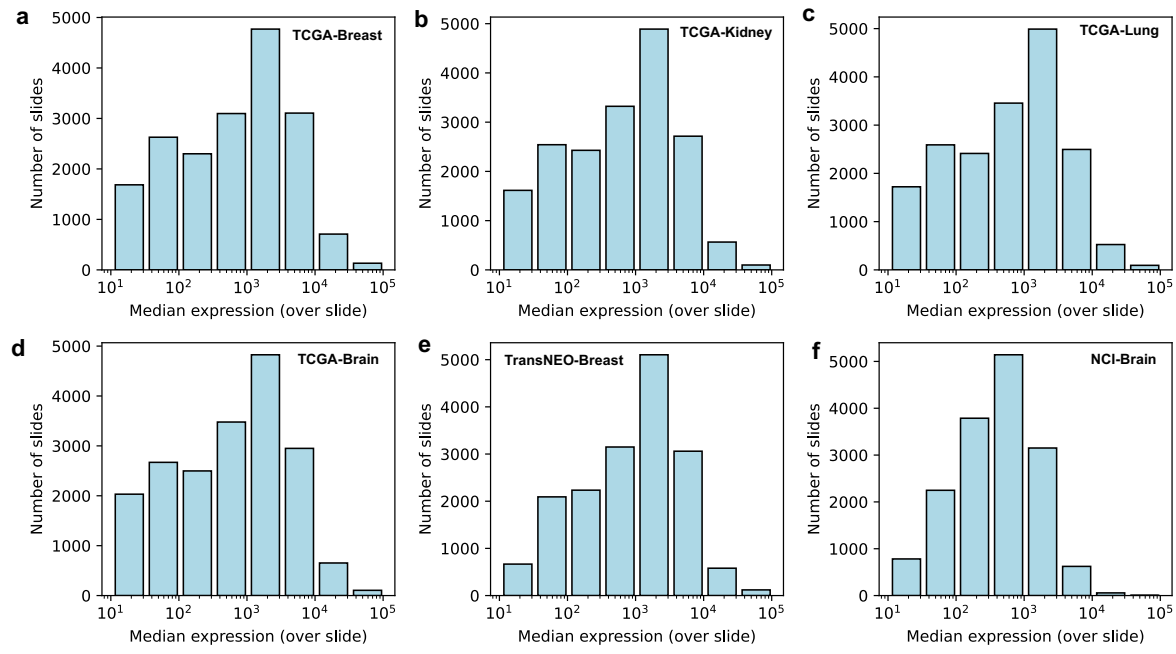
Extended Figure 8 | DeepPT performance on prediction of gene expression for the best thirty genes in the TransNEO-Breast cohort as measured by Pearson correlation (R). Notably, the models were trained on TCGA-Breast, without the need for re-training. Each scatter plot shows predicted versus actual expression for a single gene across all 160 samples.



Extended Figure 9 | DeepPT performance on prediction of gene expression for the best thirty genes in external NCI-Brain cohort as measured by Pearson correlation (R). Notably, the models were trained on TCGA-Brain cancer dataset, without the need for re-training. Each scatter plot shows predicted versus actual expression for a single gene across all 210 samples.



Extended Figure 10 | Histograms of the number of tiles per slide by cohort. The number of tiles in each slide image from TCGA and NCI-Brain datasets ranges from 100 to 8,000 (a, b, c, d), while the number of tiles in each TransNEO-Breast slide image is much smaller, ranging from 100 to 1,000 (e).



Extended Figure 11 | Histogram of median expression over slides. The median expression over samples of each gene commonly varies from 10 to 100,000 for every dataset considered in this study.